Specification-Based Software Sizing: An Empirical Investigation of Function Metrics

Ross Jeffery & John Stathis
School of Information Systems
University of New South Wales
P.O. Box 1, Kensington, NSW, 2033
AUSTRALIA

1. Introduction:

For some time the software industry has espoused the need for improved specification-based software size metrics (see Evanco et. al. 1992). During the 1980's significant resources have been applied to the development and use of metrics such as function points [Albrecht79], function weights [DeMarco82], feature points [Jones 1988] and other metrics. Earlier research [Jeffery&Low93] has established the similarity of these metrics. These metrics are used as one of the bases for cost estimation, software development management, software maintenance management, software value measurement, and so on. The proliferation of the use of the metrics and the tools now developed to support the measurement process to provide these measures, suggests that they fill an established need within the software industry. However the empirical research into these metrics has been sparse and generally not particularly favourable. Once again we see industry seeking problem solutions in the absence of experimental findings which support the solutions on offer.

This paper reports on a study of nineteen recently developed systems in a variety of application domains. The systems were developed by a single software services corporation using a variety of languages. The study investigated the following metric characteristics and questions;

Using both early and late lifecycle system documents as input to the counting process, what variation occurs in counts produced for the same system, and what gives rise to that variation? The research methodology adopted was to perform multiple independent counts of the system function size for the systems using the IFPUG Standard version 3.4. For each system this resulted in two measured function counts. The difference between these counts was analyzed both for its magnitude and the reasons for the variation. The internal validity of the function point metric was also studied and the appropriateness of the metric to the application portfolio of the organization.

This paper presents the results of this study. It is shown that:

1. Earlier research [Kitchenham 93] into inter-item correlation within the overall function count is partially supported
2. A priori function counts, in themself, do not explain the majority of the effort variation in software development in the organization studied.
3. Documentation quality is critical to accurate function identification
4. Rater error is substantial in manual function counting.

The implication of these findings for organizations using function based metrics are explored.

2. The Data Set:

The source of data for this project was an Australian software development organisation, MEGATEC Proprietary Limited, a company with approximately 50 employees that develop and distribute a range of computer software products in Australia and overseas. This organisation was selected as a test site for this work because it was one of the first software companies in Australia to gain certification to Australian Standard AS3563 for Software Quality Management. The commitment to quality in this organisation meant that managers were highly motivated to provide good quality data and there was a well established research ethic within the organisation. The 19 projects in the data set are drawn from a variety of applications. In total 17 recently completed projects were eventually included in the project database as two of the nineteen projects were not completed at the time of data analysis. A summary of the data is given in Table 1. The projects were developed during the period August 1990 to May 1993 and a high consistency in the quality staff in the use of methodology was expected in the database. The systems were written in a variety of languages including COBOL, Powerhouse, C and MS Windows, Excel Macro, SQL windows and combinations of these. It was decided that for the initial study tests would be carried out using the Albrecht Function Point counting technique as embodied in the International Function Point User Group standards as the basis for research.

TABLE I
PROJECT SIZE AND DEVELOPMENT EFFORT DATA

	Project Size (UFP)			Development Effort (Hours)		
No. Projects	Mean	Std Dev	Range	Mean	Std Dev	Range
17	551	923	38 - 3656	2093	3266	262 - 13905

Function Point were counted from documentation provided by the corporation. Each system was counted by two independent raters with experience in the IFPUG standard. One of the counters was an external consultant and the other was one of the researchers in the current study. Where we are studying the relationship between FP and other project phenomena we use the mean FP value. Data was available to derive the unadjusted function point count and also the fourteen complexity factors. In order to validate the data, structured interviews were held with all of the project managers. These interviews were used to validate the function point count, the effort data and to search for any reason behind abnormal results. There were three basic research questions which were being explored.

Firstly, we were interested in exploring in this organisational setting the relationship between development effort and function points. This question has had some considerable research over recent years, generally showing a consistent and significant relationship between the size measure and effort.

The second research question concerned replicating some of the work carried out by Kitchenham and Kansala (1993) concerning the relationships between constituent elements of the function point metric.

Thirdly, we were concerned with investigating the consistency of function point counting. There had been no study in which multiple systems were counted by multiple raters and yet it seemed that this is one of the critical elements given the current manual basis of function counting.

3. Results:

3.1 Effort Relationships

An initial Kolmogorov-Smirnov test indicated that the unadjusted and unweighted function count(UUFC), as well as the unadjusted function point (UFP) and effort data belonged to normal distributions. The results are shown in Table 2. That allowed us to proceed with a range of parametric statistical tests.

TABLE 2
KOLMOGOROV-SMIRNOV TEST

No of Projects	UUFC	UFP	Effort
	p	p	p
17	0.012	0.015	0.05

Figure 1 shows an initial plot of project size against effort for the full data set. The Project Sizing Figure 1 was unadjusted function points counted from early life-cycle documentation of a systems requirements. In this plot we can see that two of the projects are significant outliers in terms of effort and the other in terms of project size. We also note the scatter of points which has been typical in prior data when comparing size against function points. The R^2 for this data set is relatively poor showing a value of 0.228 ($p \leq 0.05$) for a linear regression of size against function points.

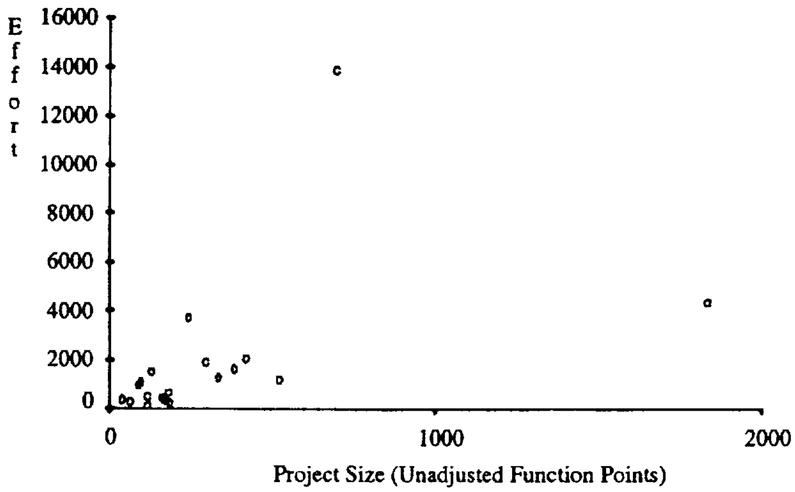


Figure 1. Scatter plot of A priori UFP against Effort

In the project manager interviews it became apparent that for some of the measured systems in the database, the project data which we show in Figure 1 was not a fair representation of the systems implemented. Taking this into account, the function point count and effort count was carried out again in order to correct any identified errors in the effort recorded or in the function point count. For example, it was found that for some of these systems the functionality had changed significantly during development and that it would be expected that a better relationship between size and effort would be found using the implemented function point count. Figure 2 shows a scatter plot for the seventeen data points after the validation of the data. The R^2 for this data set was 0.95 ($p < 0.001$). It is interesting to note the enormous difference between the data set derived at systems requirements specification stage versus the data set at implementation. This suggests that in this corporation considerable work will need to be invested to ensure requirements stability in the future if they are to gain control over predicted effort distribution.

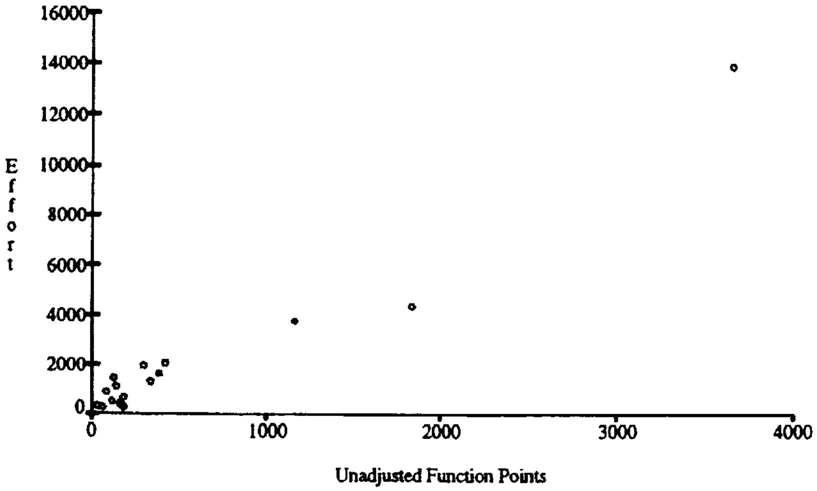


Figure 2. Scatter plot of A posteriori UFP against effort

Further analysis of the data revealed that three of the projects could be considered outliers and in line with conservative statistical analysis. Table 3 shows the regression results for the complete and the reduced data set where the outliers have been removed. Notice the reduction in the R^2 and that the effort-size relationship as expressed through the regression equation has not changed significantly suggesting that the outliers were in fact normal for this organisation.

TABLE 3
COMPARISON OF REDUCED AND FULL DATASET

	Full Dataset	Reduced Dataset
No. Projects	17	14
Equation	Effort = 192.31 + 3.45 * UFP	Effort = 187 + 4.03 * UFP
R^2	0.95 ($p<0.001$)	0.58 ($p<0.01$)

3.2 Internal Consistency of Function Points:

Table 3 shows the Pearson correlation coefficient between all pairs of function point elements using the reduced data set for conservatism. The results shows that three of the five function elements are significantly correlated. These are external inputs, external enquires and logical internal files.

TABLE 4
PEARSON CORRELATION COEFFICIENTS BETWEEN UFP ELEMENTS

Fn Point Element	**Total Unadjusted Function Point**	**EI**	**EO**	**Ext Inquiry**	**Extnl Int File**
External Input	**0.90 ($p<0.001$)**				
External Output	**0.14 (n.s.)**	**-0.07 (n.s.)**			
External Inquiry	**0.93 ($p<0.001$)**	**0.91 ($p<0.001$)**	**-0.17 (n.s.)**		
External Interface File	**-0.33 (n.s.)**	**-0.46 (n.s.)**	**0.22 (n.s.)**	**-0.45 (n.s.)**	
Logical Internal File	**0.92 ($p<0.001$)**	**0.74 ($p<0.01$)**	**-0.06 (n.s.)**	**0.90 ($p<0.001$)**	**-0.33 (n.s.)**

Kitchenham and Kansala's study used Kendall's *t* as a robust measure of correlation. In their study they found significant correlations between three pairs of function elements not reported as significant in our study. These were outputs and inputs, outputs and enquiries and outputs and internal logical files.

The results of both of these studies shows that the function elements are not independent and therefore it is possible that there may be a better effort relationship between constituent elements an effort than there is between function points. The Pearson correlation between each function point element and actual development produced the results in Table 6. These show that internal logical files and external enquiries had a higher correlation with effort than the total unadjusted function point count. This suggests that an effort estimation model derived on the internal logical file count may in fact perform better than function point for this organization.

TABLE 6
PEARSON CORRELATION RESULTS
FUNCTION ELEMENTS AGAINST EFFORT

Function Element	R^2	p
Logical Internal File	0.73	< 0.001
External Inquiry	0.63	< 0.001
External Input	0.37	< 0.001
External Output	0.03	n.s.
External Interface File	0.005	n.s.
Sum of Function Elements (UFP)	0.58	< 0.01

These results are somewhat different to Kitchenham and Kansala who found that a combination of external inputs and outputs provided a better effort predictor than unadjusted function points.

A further analysis was carried out was to compare the extent to which the complexity adjustments in the function point model add to the value of the model in explaining effort. Table 7 shows the regression results for the unadjusted and unweighted function count versus the unadjusted function point count. It can be seen from this table that once again the function point metric as a measure of size when used in its relationship with effort, appears to be performing less well than some of the constituent elements of that count.

TABLE 7
PEARSON CORRELATION RESULTS
FUNCTION ELEMENTS (UUFC & UFP) AGAINST EFFORT

	Level 1	Level 2		
	UUFC	UFP		
Function Element	R^2	p	R^2	p
Logical Internal File	**0.75**	**< 0.001**	**0.73**	**< 0.001**
External Inquiry	**0.65**	**< 0.001**	**0.63**	**< 0.001**
External Input	**0.37**	**< 0.001**	**0.37**	**< 0.001**
External Output	**0.04**	**n.s.**	**0.03**	**n.s.**
External Interface File	**0.002**	**n.s.**	**0.005**	**n.s.**
Sum of Function Elements	**0.56**	**< 0.01**	**0.58**	**< 0.01**

3.3 Rater Consistency:

The model used in this study to investigate rater consistency is shown in Figure 3 in which we see that three elements which can contribute to inconsistency. These have been identified as the system specification, the function point counting method and the rater. For example, inconsistency can be derived from the fact that the raters themselves may simply introduce errors into the function point process. It can also be that the specification can be ambiguous or at an inappropriate level of granularity such that the function point is difficult to determine, or else it could be that the function point method could be ambiguous or incomplete with respect to the function counting process that is at hand.

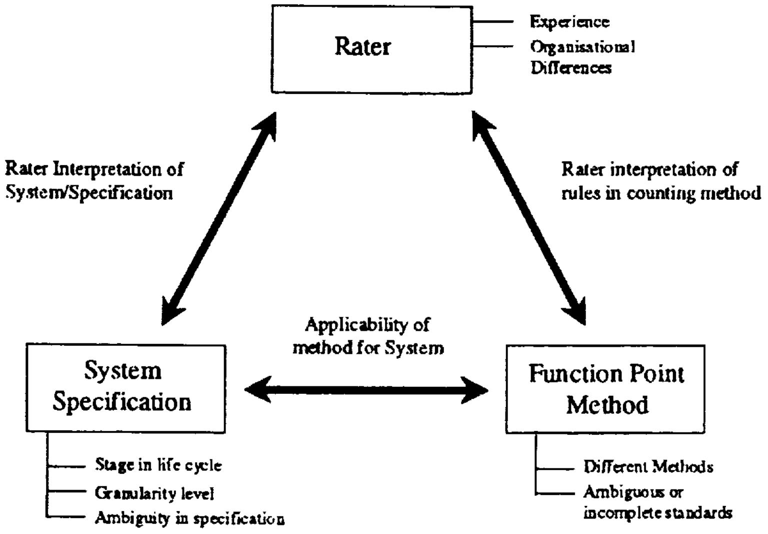

Rater
Experience
Organisational Differences
Rater Interpretation of System/Specification
Rater interpretation of rules in counting method
Applicability of method for System
System Specification
Function Point Method
Stage in life cycle
Granularity level
Ambiguity in specification
Different Methods
Ambiguous or incomplete standards

Figure 3 - A Model of the Factors Affecting Function Point Reliability

In our research we had two raters count the same systems and used variations on absolute relative difference between counts as the measure for analysis. We define the magnitude of the difference in counts between rater A and rater B as shown in equation 1 where the absolute relative is a normalised difference between the two raters normalised by average system size. We further refined this metric to the weighted absolute relative difference WARD, where we separate out the effect of each of the internal components of the function count so that errors in inputs for example, are not washed away by errors in outputs which happens if they move in opposite directions.

$$\mathrm{ARD}_{\mathrm{UFP\,(Rater\ A;\ Rater\ B)}} = \frac{\left|\mathrm{Rater\ A}_{\mathrm{UFP}} - \mathrm{Rater\ B}_{\mathrm{UFP}}\right|}{\left(\mathrm{Rater\ A}_{\mathrm{UFP}} + \mathrm{Rater\ B}_{\mathrm{UFP}}\right)\ /\ 2}$$

$$\mathrm{WARD}_{(\mathrm{EI.EO.INQ.LIF.EI;\ Rater\ A,\ Rater\ B}))} = \mathrm{ARD}_{\mathrm{EI}} \times \frac{\overline{\mathrm{EI}}_{(\mathrm{Rater\ A,\ Rater\ B})}}{\overline{\mathrm{UFP}}_{(\mathrm{Rater\ A,\ Rater\ B})}} + \ldots + \mathrm{ARD}_{\mathrm{EIF}} \times \frac{\overline{\mathrm{EIF}}_{(\mathrm{Rater\ A,\ Rater\ B})}}{\overline{\mathrm{UFP}}_{(\mathrm{Rater\ A,\ Rater\ B})}}$$

Table 8 shows the analysis results for this and in this we see that the mean WARD for these two raters is 55%. This suggests that the counting practice is relatively unstable when looked from this perspective.

Project Number	Rater A UFP	Rater B UFP	ARD	WARD	Effort	Hours Per Function Point (A, B)
Mean	**302.8**	**337.1**	**0.31**	**0.55**	**1947**	**(7.50, 6.52)**

Further analysis of this data revealed that 68% of the variation between the two counters could be attributed to rater interpretation of the specification or the application of the counting standard to that specification. Some 32% of the difference could be attributed to a simple error on the part of the rater.

4. Conclusions:

The following can be concluded from this study:

1. In a pragmatic sense the relationship between a posteriori function points and a posteriori effort is very strong for this organisation with an R^2 of .95 for the full data set or .58 for the reduced data set. This suggests that function points could be used effectively as a basis for software management in this organization.

2. From a scientific perspective it appears clear that the function point metric has some significant limitations. There is reason for concern about the function point metric. The structure of the metric is such that the components are not orthogonal which introduces issues concerning the structure of the metric. It is also of concern that the addition of the function component complexity ratings does not add to the effort relationship or the power of the effort explanation of the model. As this is counter-intuitive it warrants further investigation.

3. Inconsistency which has been observed in this study between the raters' function point counts (58%) and the high component of that difference (68%) which can be ascribed to either the function points standard or the requirements specification, suggests that the function point counting or at least the base function counting needs to be automated.

4. Given the results concerning the strong relationships between the number of internal logical files or data entities and effort, may well be possible that given further research, that if a consistent relationship holds between data entities and effort than automated size counting from data models may well be a fruitful area for further investigation.

5. References:

Albrecht,A.J. "Measuring Application Development Productivity", Share/Guide Application Development Symposium, Oct, 1979.83-92.

DeMarco, T. (1982) *Controlling Software Projects: Management, Measurement, and Estimation*, Yourdon Press, New York.

Evanco, W.M., Thomas, W.M. & Agresti W.W. "Estimating Ada System Size During Development", *Rome Laboratory Technical Report*, RL-TR-92-318, New York, December,1992.

Jeffery,D.R Low, G.& Barnes,M. "A Comparison of Function Point Counting Techniques", *IEEE Trans. on S'ware Eng.*, May 1993.

Jones,T.C. "A Short History of Function Points and Feature Points", Software Productivity Research, 1988.

Kitchenham, B & Kansala, K. "Inter Item Correlations Among Function Points", *Proc First International Software Metrics Symposium*, IEEE computer Society, Baltimore, May, 1993. 11-15.

Specification Based Software Sizing: An Empirical Investigation of Function Metrics

Ross Jeffery & John Stathis
University of New South Wales
P.OBox 1, Kensington, NSW 2033
Australia

NASA SEL Workshop 1993

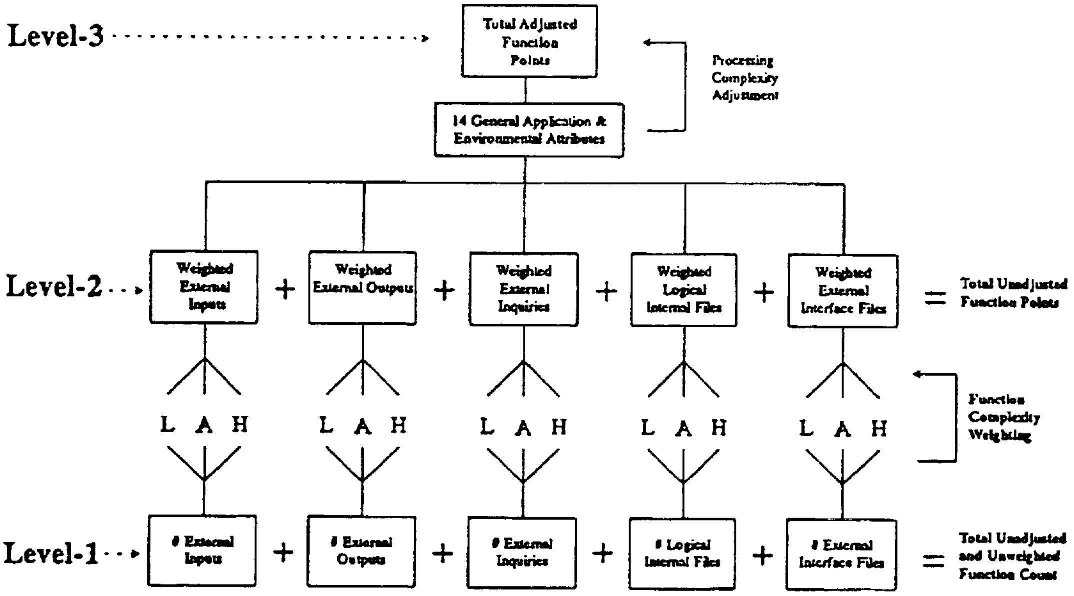
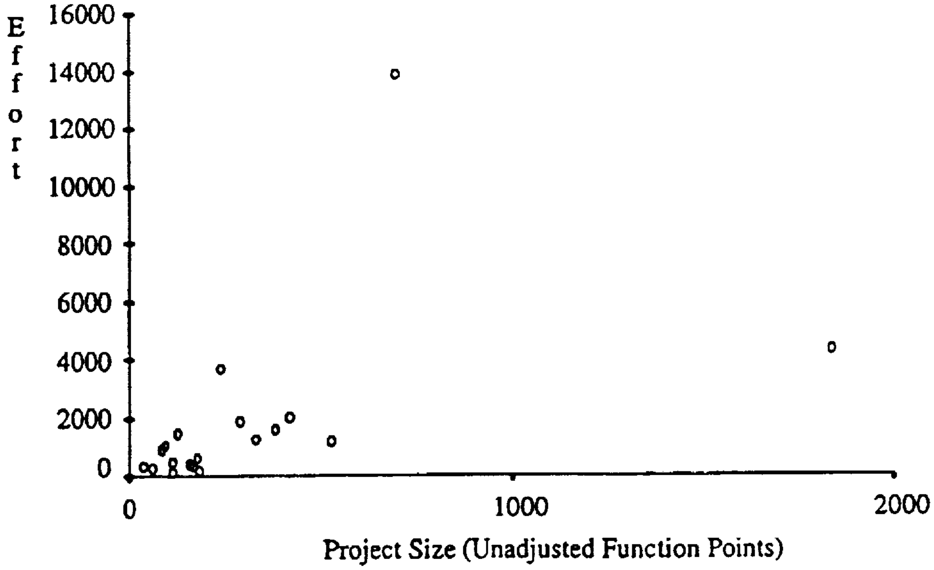

Level-3
Total Adjusted Function Points
Processing Complexity Adjustment
14 General Application & Environmental Attributes
Level-2
Weighted External Inputs
+
Weighted External Outputs
+
Weighted External Inquiries
+
Weighted Logical Internal Files
+
Weighted External Interface Files
=
Total Unadjusted Function Points
L A H
L A H
L A H
L A H
L A H
Function Complexity Weighting
Level-1
# External Inputs
+
# External Outputs
+
# External Inquiries
+
# Logical Internal Files
+
# External Interface Files
=
Total Unadjusted and Unweighted Function Count

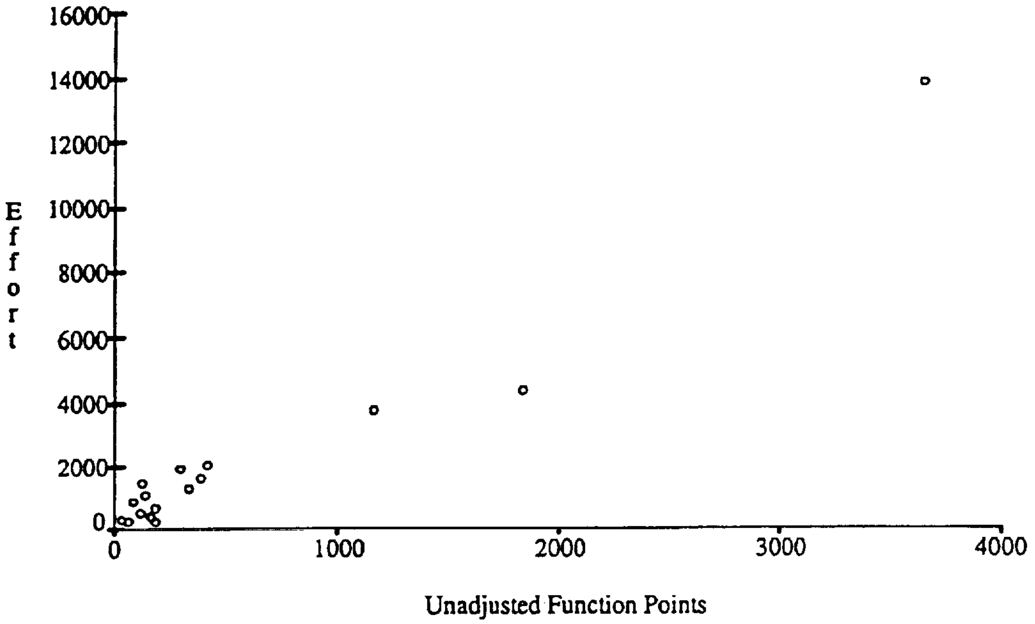
TABLE I
PROJECT SIZE AND DEVELOPMENT EFFORT DATA

	Project Size (UFP)			Development Effort (Hours)		
No. of Projects	Mean	Std. Dev.	Range	Mean	Std. Dev.	Range
17	551	923	38 - 3656	2093	3266	262 - 13905

3 *Ross Jeffery NASA SEL Workshop 1993*



4 *Ross Jeffery NASA SEL Workshop 1993*



5 *Ross Jeffery NASA SEL Workshop 1993*

TABLE II
COMPARISON OF PRE AND POST IMPLEMENTATION DATASET

	Pre Implementation FP	Post Implementation FP
No. Projects	**17**	**17**
Regression Equation (UFP against effort)	**effort = 914.6 + 3.7 * UFP**	
R^2 (p)	**0.228 (0.05)**	**0.95**

6 *Ross Jeffery NASA SEL Workshop 1993*

TABLE III
COMPARISON OF REDUCED AND FULL DATASET

	Full Dataset	Reduced Dataset
No. Projects	17	14
Regression Equation (UFP against effort)	effort = 192.31 + 3.45 * UFP	effort = 185.37 + 4.03 * UFP
R^2 (p)	0.95 ($p < 0.001$)	0.58 ($p < 0.01$)

7 *Ross Jeffery NASA SEL Workshop 1993*

TABLE IV
PREVIOUS STUDIES - UFP AGAINST EFFORT

Study	No. of Projects	Unadjusted Function Points	
		R^2	(p)
Albrecht and Gaffney, 1983	24	0.90	< 0.001
Kemerer, 1987	15	0.54	< 0.001
Kitchenham and Kansala, 1993	40	0.41	< 0.01
Jeffery et. al., 1993	64	0.36	< 0.001
Jeffery & Stathis, Current Study	14	0.58	< 0.001

8 *Ross Jeffery NASA SEL Workshop 1993*

TABLE V
PEARSON CORRELATION COEFFICIENTS BETWEEN UFP ELEMENTS

Function Point Element	Total Unadjusted Function Point	External Input	External Output	External Inquiry	External Interface File
External Input	0.90 (p<0.001)				
External Output	0.14 (n.s.)	-0.07 (n.s.)			
External Inquiry	0.93 (p<0.001)	0.91 (p<0.001)	-0.17 (n.s.)		
External Interface File	-0.33 (n.s.)	-0.46 (n.s.)	0.22 (n.s.)	-0.45 (n.s.)	
Logical Internal File	0.92 (p<0.001)	0.74 (p<0.01)	-0.06 (n.s.)	0.90 (p<0.001)	-0.33 (n.s.)

TABLE VI
PEARSON CORRELATION RESULTS
FUNCTION ELEMENTS AGAINST EFFORT

Function Element	R^2	p
Logical Internal File	0.73	< 0.001
External Inquiry	0.63	< 0.001
External Input	0.37	< 0.001
External Output	0.03	n.s.
External Interface File	0.005	n.s.
Sum of Function Elements (UFP)	0.58	< 0.01

TABLE VII
PEARSON CORRELATION RESULTS
FUNCTION ELEMENTS (UUFC & UFP) AGAINST EFFORT

	Level 1		Level 2	
	UUFC		UFP	
Function Element	R^2	p	R^2	p
Logical Internal File	0.75	< 0.001	0.73	< 0.001
External Inquiry	0.65	< 0.001	0.63	< 0.001
External Input	0.37	< 0.001	0.37	< 0.001
External Output	0.04	n.s.	0.03	n.s.
External Interface File	0.002	n.s.	0.005	n.s.
Sum of Function Elements	0.56	< 0.01	0.58	< 0.01

11 *Ross Jeffery NASA SEL Workshop 1993*

TABLE VIII
EFFORT ESTIMATE ARE t-TESTS FOR
UUFC AND UFP

	Unweighted and Unadjusted Function Count (UUFC)		Unadjusted Function Point (UFP)			
No. of Projects	**Mean ARE**	**Std. Dev.**	**Mean ARE**	**Std. Dev.**	*t*	*p*
17	0.53	0.64	0.51	0.60	0.70	0.492

12 *Ross Jeffery NASA SEL Workshop 1993*

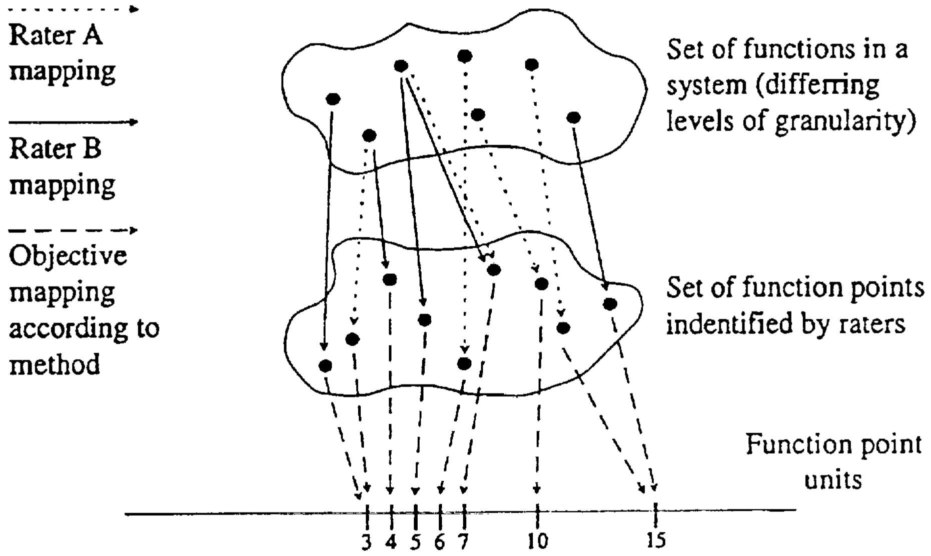

Rater A mapping
Rater B mapping
Objective mapping according to method
Set of functions in a system (differring levels of granularity)
Set of function points indentified by raters
Function point units
3
4
5
6
7
10
15

Figure II - Mapping a Set of Functions to Function Point Units

13 *Ross Jeffery NASA SEL Workshop 1993*

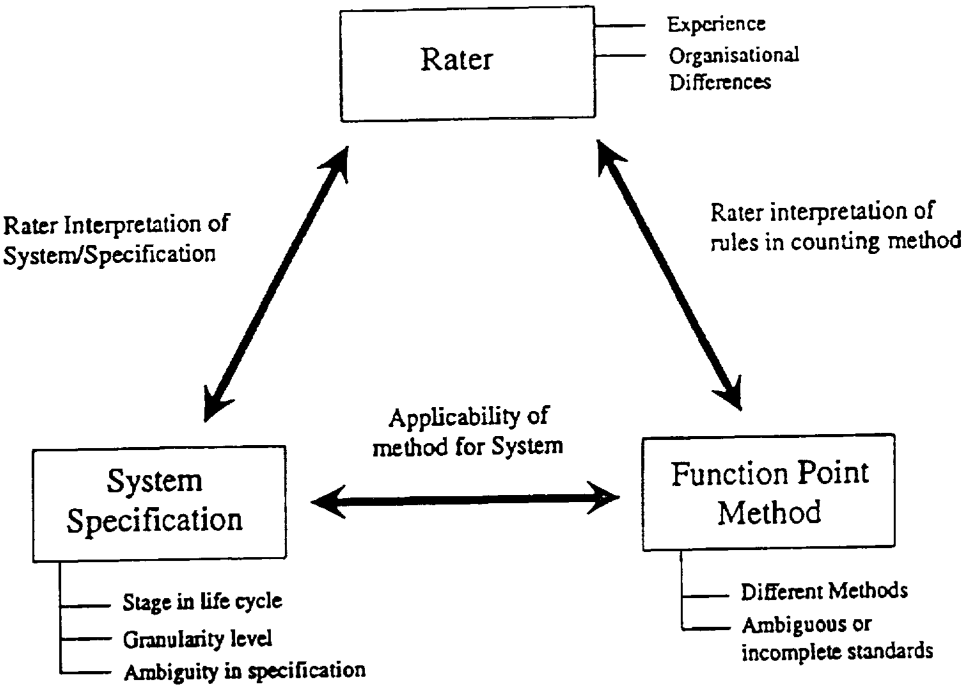

Rater
Experience
Organisational Differences
Rater Interpretation of System/Specification
Rater interpretation of rules in counting method
Applicability of method for System
System Specification
Function Point Method
Stage in life cycle
Granularity level
Ambiguity in specification
Different Methods
Ambiguous or incomplete standards

Figure III - A Model of the Factors Affecting Function Point Reliability

14 *Ross Jeffery NASA SEL Workshop 1993*

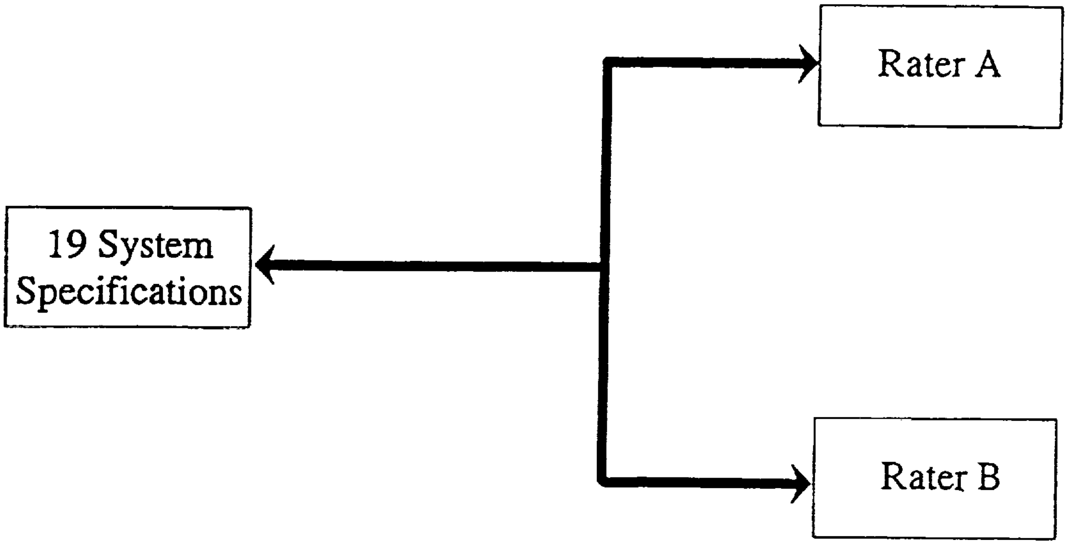


Figure VIII - Research Design for Current Study

15 *Ross Jeffery NASA SEL Workshop 1993*

$$ARD_{UFP\ (Rater\ A;\ Rater\ B)} = \frac{|Rater\ A_{UFP}\ -\ Rater\ B_{UFP}|}{(Rater\ A_{UFP}\ +\ Rater\ B_{UFP})\ /\ 2}$$

$$= 31\%$$

$$WARD_{(EI.EO.INQ.LIF.EIF;\ Rater\ A,\ Rater\ B))}$$

$$= ARD_{EI} \times \frac{\overline{EI}_{(Rater\ A,\ Rater\ B)}}{\overline{UFP}_{(Rater\ A,\ Rater\ B)}} + \ldots$$

$$+ ARD_{EIF} \times \frac{\overline{EIF}_{(Rater\ A,\ Rater\ B)}}{\overline{UFP}_{(Rater\ A,\ Rater\ B)}}$$

$$= 55\%$$

16 *Ross Jeffery NASA SEL Workshop 1993*

MEAN ABSOLUTE RELATIVE DIFFERENCE (MARD) UNWEIGHTED AND WEIGHTED FUNCTION POINTS

	Total Function Point Count (UUFC) (UFP)	External Input	External Output	External Inquiry	External Interface File	Logical Internal File
Unweighted Function Points	0.33	0.76	0.69	0.65	0.54	0.45
Weighted Function Points	0.31	0.67	0.70	0.62	0.54	0.43

1. Strong a posteriori function points and a posteriori effort relationship for this organisation - R^2 of 0.95 for the full data set or 0.58 for the reduced data set.

2. The function point metric has some significant limitations.

   Components are not orthogonal

   Function component complexity ratings does not add to the effort explanation of the model.

3. Inconsistency has been observed between the raters' function point counts (58%)

   A high component of that difference (68%) can be ascribed to either the function points standard or the requirements specification

4. Automated size counting from data models may well be a fruitful area for further investigation.